CONCENTRATION INEQUALITIES

1. AZUMA-HOEFFDING INEQUALITY

Chebyshev's inequality is one of the most useful gadgets in the toolbox of probability, but it does have its drawbacks. For estimating probabilities of somewhat rare events, e.g., .05 or less, it is often far too conservative, and so it isn't so good for setting confidence intervals in estimation problems. Furthermore, as $\alpha \to \infty$ the Chebyshev bound on $P\{S_n > \alpha\}$ decays only like C/α^2, even though the probabilities themselves might actually decay *exponentially* in α. The subject of *concentration inequalities* is about finding improvements of various sorts to Chebyshev. One of the most useful concentration inequalities is the *Azuma-Hoeffding* inequality for sums of *bounded* random variables.

Theorem 1. *(Azuma-Hoeffding) Let S_n be a martingale (relative to some sequence $Y_0, Y_1, \dots$) satisfying $S_0 = 0$ whose increments $\xi_n = S_n - S_{n-1}$ are bounded in absolute value by 1. Then for any $\alpha > 0$ and $n \geq 1$,*

$$P\{S_n \geq \alpha\} \leq \exp\{-\alpha^2/2n\}. \tag{1}$$

More generally, assume that the martingale differences ξ_k satisfy $|\xi_k| \leq \sigma_k$. Then

$$P\{S_n \geq \alpha\} \leq \exp\left\{-\alpha^2/2\sum_{j=1}^{n}\sigma_j^2\right\}. \tag{2}$$

In both cases the denominator in the exponential is the maximum possible variance of S_n subject to the constraints $|\xi_n| \leq \sigma_n$, which suggests that the worst case is when the distributions of the increments ξ_n are as spread out as possible. The following lemma suggests why this should be so.

Lemma 1. *Among all probability distributions on the unit interval $[0,1]$ with mean p, the most spread-out is the Bernoulli-p distribution. In particular, for any probability distribution F on $[0,1]$ with mean p and any convex function $\varphi : [0,1] \to \mathbb{R}$,*

$$\int_0^1 \varphi(x)\, dF(x) \leq p\varphi(1) + (1-p)\varphi(0).$$

Proof. This is just another form of Jensen's inequality. Let X be a random variable with distribution F, and let U be an independent uniform-[0,1] random variable. The indicator $1\{U \leq X\}$ is a Bernoulli r.v. with conditional mean

$$E(1\{U \leq X\} \,|\, X) = X,$$

and so by hypothesis the *unconditional* mean is $EE(1\{U \leq X\} \,|\, X) = EX = p$. Thus, $1\{U \leq X\}$ is Bernoulli-p. The conditional version of Jensen's inequality implies that

$$E(\varphi(1\{U \leq X\}) \,|\, X) \geq \varphi(E(1\{U \leq X\} \,|\, X)) = \varphi(X).$$

Taking unconditional expectation shows that

$$E\varphi(1\{U \le X\}) \ge E\varphi(X),$$

which, since $1\{U \le X\}$ is Bernoulli-p, is equivalent to the assertion of the lemma. □

Rescaling gives the following consequence (exercise).

Corollary 1. *Among all probability distributions on the interval $[-A,B]$ with mean zero, the most spread out is the two-point distribution concentrated on $-A$ and B that has mean zero. In particular, if φ is convex on $[-A,B]$ then for any random variable X satisfying $EX = 0$ and $-A \le X \le B$,*

$$E\varphi(X) \le \varphi(-A)]\frac{B}{A+B} + \varphi(B)\frac{A}{A+B}. \tag{3}$$

In particular, if $A = B > 0$ then for any $\theta > 0$,

$$Ee^{\theta X} \le \cosh(\theta A) \tag{4}$$

Proof. The first statement follows from Lemma 1 by rescaling, and the cosh bound in (**??**) is just the special case $\varphi(x) = e^{\theta x}$. □

Lemma 2. $\cosh x \le e^{x^2/2}$.

Proof. The power series for $2\cosh x$ can be gotten by adding the power series for e^x and e^{-x}. The odd terms cancel, but the even terms agree, so

$$\cosh x = \sum_{n=0}^{\infty} \frac{x^{2n}}{(2n)!} \le \sum_{n=0}^{\infty} \frac{x^{2n}}{2^n(n!)} = \exp\{x^2/2\}.$$

□

Conditional versions of Lemma 1 and Corollary **??** can be proved by virtually the same arguments (which you should fill in). Here is the conditional version of the second inequality in Corollary **??**.

Corollary 2. *Let X be a random variable satisfying $-A \le X \le A$ and $E(X \mid Y) = 0$ for some random variable or random vector Y. Then for any $\theta > 0$,*

$$E(e^{\theta X} \mid Y) \le \cosh(\theta A) \le e^{\theta^2 A^2/2}. \tag{5}$$

Proof of Theorem 1. The first inequality (**??**) is obviously a special case of the second, so it suffices to prove (**??**). By the Markov inequality, for any $\theta > 0$ and $\alpha > 0$,

$$P\{S_n \ge \alpha\} \le \frac{Ee^{\theta S_n}}{e^{\theta\alpha}}. \tag{6}$$

Since $e^{\theta x}$ is a convex function of x, the expectation on the right side can be bounded using Corollary **??**, together with the hypothesis that $E(\xi_k \mid \mathscr{F}_{k-1}) = 0$. (As usual, the notation $\mathscr{F}_k$ is

just shorthand for conditioning on $Y_0, Y_1, Y_2, \ldots, Y_k$.) The result is

$$\begin{aligned} Ee^{\theta S_n} &= EE(e^{\theta S_n} \mid \mathscr{F}_{n-1}) \\ &\le Ee^{\theta S_{n-1}} E(e^{\theta \xi_n} \mid \mathscr{F}_{n-1}) \\ &\le Ee^{\theta S_{n-1}} \cosh(\theta \sigma_n) \\ &\le Ee^{\theta S_{n-1}} \exp\{\theta^2 \sigma_n^2/2\} \end{aligned} \tag{7}$$

Now the same procedure can be used to bound $Ee^{\theta S_{n-1}}$, and so on, until we finally obtain

$$Ee^{\theta S_n} \le \prod_{k=1}^{n} \exp\{\theta^2 \sigma_n^2/2\}$$

Thus,

$$P\{S_n \ge \alpha\} \le e^{-\theta\alpha} \prod_{k=1}^{n} \exp\{\theta^2 \sigma_n^2/2\}$$

for *every* value $\theta > 0$. A sharp inequality can now be obtained by choosing the value of θ that minimizes the right side, or at least a value of θ near the min. A bit of calculus shows that the minimum occurs at

$$\theta = \alpha / \sum_{k=1}^{n} \sigma_k^2.$$

With this value of θ, the bound becomes

$$P\{S_n \ge \alpha\} \le \exp\left\{-\alpha^2/2 \sum_{j=1}^{n} \sigma_j^2\right\}.$$

□

2. McDiarmid's Inequality

Theorem 1 was first discovered by Hoeffding and then again independently by Azuma. Both had different applications in mind, but both were mainly concerned with sums of uncorrelated random variables. In the late 1980s a number of researchers (Talagrand, McDiarmid, and others) began to realize that the Azuma-Hoeffding inequality had important implications for *nonlinear* functions of bounded random variables.

Theorem 2. *(McDiarmid) Let $X_1, X_2, \ldots, X_n$ be independent random variables such that $X_i \in A_i$, for some Borel sets A_i. Suppose that $f : \prod_{i=1}^n A_i \to \mathbb{R}$ is "Lipschitz" in the following sense: for each $k \le n$ and any two sequences $x, x' \in \prod_{i=1}^n A_i$ that differ only in the kth coordinate,*

$$|f(x) - f(x')| \le \sigma_k. \tag{8}$$

Let $Y = f(X_1, X_2, \ldots, X_n)$. Then for any $\alpha > 0$,

$$P\{|Y - EY| \ge \alpha\} \le 2\exp\{-2\alpha^2 / \sum_{k=1}^{n} \sigma_k^2\}. \tag{9}$$

Proof. We will want to condition on the first k of the random variables X_i, so we will denote by $\mathscr{F}_k$ the $\sigma-$algebra generated by these r.v.s. Let

$$Y_k = E(Y \mid \mathscr{F}_k) = E(Y \mid X_1, X_2, \ldots, X_k).$$

Then the sequence X_k is a martingale (why?). Furthermore, the successive differences satisfy

$$|Y_k - Y_{k-1}| \leq \sigma_k.$$

To see this, let $Y' = f(X')$, where X' is obtained from X by replacing the kth coordinate X_k with an independent copy X'_k and leaving all of the other coordinates alone. Then

$$E(Y' \mid \mathscr{F}_k) = E(Y \mid \mathscr{F}_{k-1}) = Y_{k-1}$$

But by hypothesis, $|Y' - Y| \leq \sigma_k$. This implies that

$$|Y_k - Y_{k-1}| = |E(Y - Y' \mid \mathscr{F}_k)| \leq \sigma_k.$$

Given this, the result follows directly from the Azuma-Hoeffding inequality, because $Y = E(Y \mid \mathscr{F}_n)$ and $EY = E(Y \mid \mathscr{F}_0)$.

□